Who's Harry Potter? Approximate Unlearning in LLMs

Year: 2023 | Citations: 303 | Authors: Ronen Eldan, M. Russinovich

Abstract

Large language models (LLMs) are trained on massive internet corpora that often contain copyrighted content. This poses legal and ethical challenges for the developers and users of these models, as well as the original authors and publishers. In this paper, we propose a novel technique for unlearning a subset of the training data from a LLM, without having to retrain it from scratch. We evaluate our technique on the task of unlearning the Harry Potter books from the Llama2-7b model (a generative language model recently open-sourced by Meta). While the model took over 184K GPU-hours to pretrain, we show that in about 1 GPU hour of finetuning, we effectively erase the model's ability to generate or recall Harry Potter-related content, while its performance on common benchmarks (such as Winogrande, Hellaswag, arc, boolq and piqa) remains almost unaffected. We make our fine-tuned model publicly available on HuggingFace for community evaluation. To the best of our knowledge, this is the first paper to present an effective technique for unlearning in generative language models. Our technique consists of three main components: First, we use a reinforced model that is further trained on the target data to identify the tokens that are most related to the unlearning target, by comparing its logits with those of a baseline model. Second, we replace idiosyncratic expressions in the target data with generic counterparts, and leverage the model's own predictions to generate alternative labels for every token. These labels aim to approximate the next-token predictions of a model that has not been trained on the target data. Third, we finetune the model on these alternative labels, which effectively erases the original text from the model's memory whenever it is prompted with its context.